Machine Learning Lunch Series Problem 3

Thomas Britton (tbritton@jlab.org) David Lawrence (davidl@jlab.org)

Nov 2019

1 Introduction

1.1 Setup

Arguably, the most important part of machine learning is understanding why the model makes the decision(s) it does. Understanding the "whys" can also yield increases in training efficiency as a more optimal and tailored set, containing only the "important" examples can be found. In this vein there exist, for models, so called "adversarial examples". This is typically shown through the befuddlement of a classifier through subtle modification of an image. These examples fool models in often hilarious ways while being unable to stump a human. For example, in Figure 1 a classifier has confused this picture of a cat with a picture of guacamole. Recently, a turtle has been 3D printed which fools a google classifier from a multitude of angles (Fig 2).

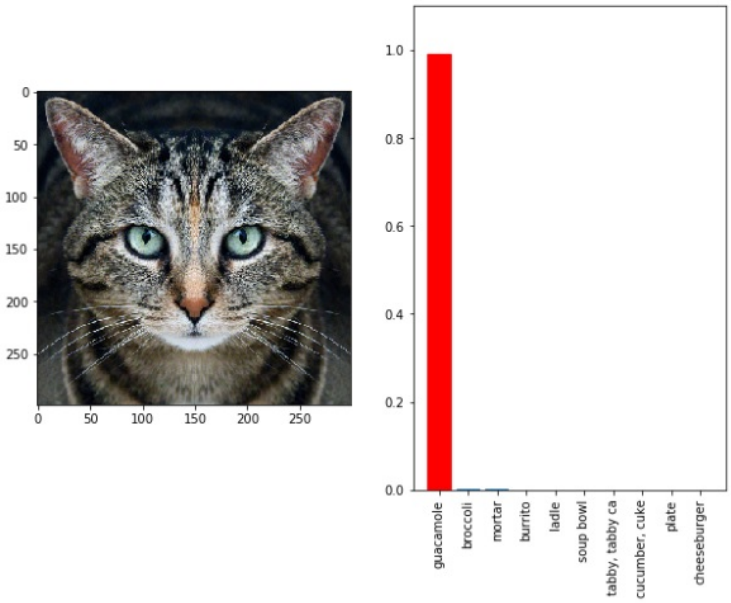

Figure 1: MIT's LabSix fools Inception v3 into classifying a cat as guacamole.**Fooling Neural Networks in the Physical World with 3D Adversarial Objects**[1])

1.2 Goal

The goal will be to perform a "one pixel attack" on a network trained to classify detector occupancy as being either good or bad. Note: this attack need not be confined to a single pixel. Contestants will be given a target image to attack with the goal of confusing a pre-trained model through minimal image modification.

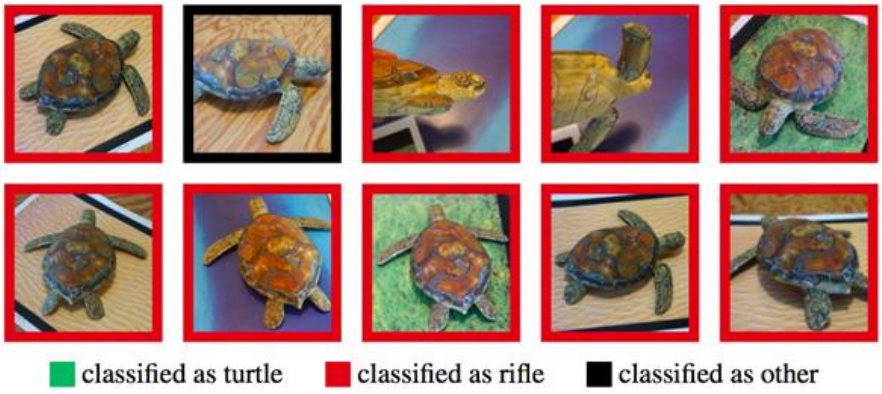

Figure 2: MIT plastic "adversarial object" confuses Google's Inception v3 Classifier. (**Synthesizing Robust Adversarial Examples**[2])

2 Materials

The given materials will be a pre-trained model (tested compatible with python 3.6.8, tensorflow 1.14.0, and Keras 2.3.1), a target image named "TARGET_IMG.png", an inference script, as well as an example image directory with examples of images from each label ("Good","Bad","NoData"). Note: The model/script may be compatible with other versions of the various software packages, though we only guarantee compatibility with the above versions. To aid contestants an example target has also been given (EXAMPLE_OF_THE_TARGET_IMG.png) along with a corresponding example submission (EXAMPLE_SUBMISSION_AGAINST_EXAMPLE_TARGET.png). This example submission does indeed change the classification of EXAMPLE_OF_THE_TARGET_IMG.png from "Bad" to "Good" and thus constitutes a valid submission. It, however, does not fool a person and thus represents a fairly week adversarial example.
The script takes in one argument (-D [path/to/IMG.png]). Inclusive of load times for the tensorflow libraries and model it takes a little over 5 seconds to run. At the end of running 4 lines will be printed above the "Processing time" line (e.g.):
[0]
{'Bad': 0, 'Good': 1, 'NoData': 2}
[[9.9998045e-01 1.9142049e-05 3.0581936e-07]]
Bad

The first of these lines is a python array containing the predicted index (0 in this case). The second line is a dump of the dictionary mapping the above predicted index to an index. The third line is an array of arrays of confidences in each label. Because this script is set up to infer on a single image it is just an array with a single array element containing the confidences. It is important to note that sum of confidences is approximately unity. The last of the 4 lines is the plain-text label inferred. In this case the inferred image has been labeled "Bad by the model with a confidence of 9.9998045e-01. All materials will be available as usual **here**[3]. Independent copies can be requested by emailing tbritton@jlab.org.

3 Judging Criteria

Contestants need only turn in a (the) modified color image in png format, this image should be of the same size as the target image. This image will then be fed to the model provided and a difference from original image calculated. The winner will be the individual(s) which most fool the model with the least change to the provided target image. In the event no one is able to fool the machine the winner will be the one who most harms the model's confidence in the classification of the unaltered target image

[3]https://halldweb.jlab.org/talks/ML_lunch/Nov2019/

(currently 99.998045% confidence the target image is "Bad"). Submissions are due Feb 5th 2020 by noon est.

4 Prizes

Prizes will be mainly in the form of glory and (local) fame. A couple of gift cards will also be thrown in.